中国电子文件知识组织XML集成置标标准化研究

段荣婷 著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

中国电子文件知识组织
XML 集成置标标准化研究

段荣婷 著

上海交通大學出版社

内 容 提 要

本书分两部分。理论部分，对电子文件集成置标的概念、意义、基本性质、基本结构、功能、原理、体系的静态与动态构建、基础语言以及电子文件文本需求体系构建等进行了系统研究。应用部分，较为系统地探讨了处于电子文件生命周期各阶段的中国电子文件置标典型案例，具体包括：基于可扩展置标语言(XML)的、面向文件的中国电子公文文本置标；基于资源描述框架(RDF)的、面向数据的中国电子文件背景信息文本置标；基于 RDF 的、面向数据的中国电子文件机读目录文本置标；基于简约知识组织系统(SKOS)的、面向数据的《中国档案主题词表》文本置标。全书图文并茂、紧跟国际学术前沿、创新性与实用性强，强调基于 XML 的中国电子文件集成置标的实用性和可操作性。

本书适合于档案工作者、文件管理者以及 XML 爱好者阅读，也可供图书情报档案专业的教师、学生和科研人员参考。

图书在版编目(CIP)数据

中国电子文件知识组织 XML 集成置标标准化研究/段荣婷著. —上海：上海交通大学出版社，2012

ISBN 978-7-313-07681-6

Ⅰ. 中… Ⅱ. 段… Ⅲ. ①电子文件—档案管理—研究 Ⅳ. G276

中国版本图书馆 CIP 数据核字(2011)第 169556 号

中国电子文件知识组织
XML 集成置标标准化研究

段荣婷 著

上海交通大学出版社出版发行

(上海市番禺路 951 号 邮政编码 200030)

电话：64071208 出版人：韩建民

浙江云广印业有限公司 印刷 全国新华书店经销

开本：787mm×960mm 1/16 印张：30 字数：567 千字

2012 年 5 月第 1 版 2012 年 5 月第 1 次印刷

ISBN 978-7-313-07681-6/G 定价：98.00 元

前　言

“电子文件置标”既是电子文件管理在现有人-机系统环境下迫切需要解决的重大现实性课题，又是电子文件未来面临基于信息网格与云计算这种新型人-机系统环境必须解决的前瞻性课题，也是在电子文件管理中，确保电子文件长久保存、确保其凭证与情报价值实现的重要而专门的研究方向。

本书的研究目的在于透过多维集成置标的视角，从理论与应用两个层面探索电子文件凭证性和长久保存的实现方式，进而为电子文件在其整个生命周期中的科学置标、科学管理提供理论与实践支撑。

本书的主要创新是提出了电子文件集成置标的全新视角，即以电子文件为核心，从时间维度的电子文件生命周期视角出发，以确保电子文件凭证价值得到长期、永久的保存、共享与利用；从空间维度的文本视角、置标所用语言视角、人-机环境视角出发，对电子文件置标的理论与实践进行多维、立体、综合的系统研究。这种研究方法的优点是既能确保研究的全面性，又有利于研究的深入性。

1. 主要内容

基于多维及理论与应用两层面集成置标的研究思路，本书采用递进式逻辑结构体系。

1）理论部分

第 1 章为导论。论述了研究的必要性，并进行了可行性分析，明确了论文研究设计思路、研究内容结构、研究方法和研究的创新之处。

第 2 章为国际电子文件置标理论与应用。首先综述目前国内外对本选题及相关课题的研究现状和已取得的研究成果，说明本选题的重要学术与技术应用价值；然后拓宽研究的视野，力求将本研究建立在比较高的起点上。

第 3 章为电子文件置标基本理论。重点揭示了电子文件置标的概念内涵与主要特征等基本理论，这是论文的逻辑起点。

第 4 章为基于 XML 的电子文件置标体系与基础置标用语言的专门理论. 首先论述了电子文件置标的静态与动态体系构建及其多维置标原则；其次论述了电子

文件优选基础置标用语言——可扩展置标语言(XML)及其置标用模式语言(XML Schema)。

第5章为基于XML的电子文件置标需求理论,对置标需求体系的构建进行了全面深入的分析。

2) 应用部分

第6章为基于XML的面向文件的中国电子公文文本置标。

第7章为基于RDF的面向数据的中国电子文件背景信息文本置标。

第8章为基于RDF的面向数据的中国电子文件机读目录文本置标。

第9章为基于SKOS的面向数据的电子文件概念体系文本置标。

第6-9章分别探讨了中国电子文件公文文本、中国电子文件背景信息文本、中国电子文件机读目录文本和中国电子文件概念体系文本置标的技术体系结构与内容等。

第10章为基于XSL的中国电子文件集成可视化置标应用及检索效率评价。

本书理论部分的第3、4与5章,及应用部分,是核心部分;最后是结论及有待进一步研究的问题。

2. 主要特点

本书全面介绍了XML技术开发和应用知识,具有以下特点:

1) 内容系统

本书是以中国电子文件为核心,从时间维度的电子文件生命周期视角出发,以确保电子文件凭证价值得到长期、永久的保存、共享与利用;从空间维度的文本视角、置标所用语言视角、人—机环境视角出发,对中国电子文件置标进行多维、综合的系统研究,即不仅要研究中国电子文件的逻辑置标、附含语义置标,而且还要研究本体置标;不仅要研究可扩展置标语言(XML)置标,而且还要综合研究资源描述框架(RDF)、简约知识组织体系(SKOS)本体语言置标;不仅要应用前端置标、中端置标,而且还要应用后端置标,乃至全程置标。集成置标是系统研究,即不仅是不同视角的多维研究,亦是不同层次的立体研究,即在理论与应用这两个层次上对中国电子文件置标作出整体、全面的深入研究。

2) 创新性强

创新特色体现于特点新、角度新、思路新及方法新等方面。理论创新与实践创新均构成一个整体,理论创新的整体性突出表现为基于XML的中国电子文件集成置标研究的深度、广度与前瞻性和新颖性;实践创新体现了置标实现的动态逻辑过程,其整体性突出表现为基于XML的中国电子文件集成置标研究的可行性、应用性与科学的最佳实践性和与时俱进的创新性。

3）简明实用

本书所述逐级深化的全程集成置标，就单份电子文件而言，有利于实现电子文件的全程控制；而就整体而言，有利于实现电子文件全宗—类别—案卷—文件的多级置标，确保全宗来源原则的实现和全宗电子文件的完整齐全，最终确保电子文件的永久保存及其凭证与情报价值的实现，这对于我国现今科学构建覆盖人民群众的电子文件资源体系、方便人民群众的电子文件利用体系和确保电子文件安全保密的电子文件安全体系，具有十分重大的意义。

目　录

理　论　篇

应 用 篇

理　论　篇

第1章 导 论

1.1 研究背景及研究问题的提出

所谓的研究背景，即本书选题所涉及的时代背景与社会背景。时代背景主要是基于时间轴上的电子文件及其管理的发展对电子文件置标所形成的背景；社会背景主要是空间轴上的在文件/档案领域中电子文件管理的研究与应用状况。

1.1.1 时代背景：信息时代电子文件作为数字遗产保存的紧迫性

自1946年世界上第一台电子计算机在美国诞生，电子文件就开始步入人类社会，并成为人类共同的数字遗产。随着计算机、网络等信息技术的迅猛发展，信息时代的电子文件迅速增长与普及。电子文件越来越广泛和深入地渗透和影响着人类社会生活的几乎所有领域，其快速增长的庞大数量，以及对社会生活真切反映和普遍联系所造就的内在质量，使得电子文件成为现代社会信息资源的重要组成部分。因而，电子文件管理也越来越受到各国政府和公众的重视，成为国际档案界最关注的焦点之一，也成为档案信息化与数字档案馆等建设的重要组成部分。

同样在我国，电子文件数量也迅速增长。据调查，目前大部分单位电子文件已经占其全部文件数量的50%以上，电子文件已经成为业务活动的主要记录形式。

但是，电子文件管理仍面临丢失、凭证效用无法确保等风险[①]。归纳起来，电子文件管理中存在的问题主要包括：①电子文件格式不规范，无法满足交换、共享与长久保存、利用的需求；②电子文件元数据管理的互操作性差，从而影响了电子文件的真实性、完整性和可用性，电子文件作为战略资源的功能与凭证及情报价值都无法发挥。

2004 年联合国教科文组织在其发布的《保存数字遗产宪章》中指出：数字化遗产是共同遗产。它们存在的时间一般不长，需要有意地制作、维护和管理才能保存下来。这类资源大多具有长久的价值和意义，因而是一种应为当代人和后代人加以保护和保存的遗产。各种语言、世界各地和人类的各种知识或表达方式都可能有这种呈增长趋势的遗产。世界上的数字遗产面临着消失和失传的危险，如果不着手解决目前所面临的有关威胁，数字遗产将会迅速丢失，而且不可避免[②]。

同样，作为人类共同的数字遗产的重要组成部分——电子文件，如果不着手解决其凭证价值所面临的有关威胁，电子文件的丢失，也将是不可避免的。

1.1.2 社会背景：国际文件/档案领域对电子文件管理的高度关注

从文件/档案领域出发，要保存电子文件，最主要就是使电子文件在空间上与时间上实现跨平台互操作，即要实现电子文件长久保存。因此，电子文件长久保存、有效利用不仅仅是现在，而且也是未来长久开发档案信息资源的战略步骤，是保护人类数字遗产的重大战略措施。正是由于意识到了这一点，国际文件/档案领域从 20 世纪 90 年代就开始了对电子文件管理的探索，到 21 世纪初更是得到长足

---

① 作者注：据中国档案报 2008 年 3 月 13 日第 003 版专题报道，中国档案学会和中国人民大学信息资源管理学院联合组成“电子文件管理机制研究”课题组。2007 年，接受课题组抽样调查的 55 家中央机关及其直属企事业单位生成的电子文件数量已经占其全部文件数量的 72.7%。其中，42.2%的电子文件没有以任何方式有效留存。74.4%的单位没有采用任何措施存留数据库文件、电子邮件、音频文件、视频文件、多媒体文件、超媒体文件、网页文件等类型的电子文件。一些留存下来的电子文件因管理不善而无法读取。过去三年间，已有 22.5%的中央单位不同程度地出现过电子文件不可读现象。73.6%的中央单位还承认，因为相关法规制度不健全、电子文件元数据不完整等原因，其自身生成的电子文件无法独立发挥文件的功效；在接受抽样调查的 35 家省级、副省级城市国家综合档案馆中，86.2%的档案馆保存的电子文件不具有证据效力。

② United Nations Educational, Scientific and Cultural Organization. Charter on the Preservation of the Digital Heritage. Adopted at the 32nd session of the General Conference of UNESCO, 17 October 2003: 1-2.

发展。

1999年,国际上最有影响的电子文件管理研究项目之一"电子系统中电子文件真实性永久保存的国际研究项目[①](简称 InterPARES)"开始启动。该项目从1999年至2001年完成了第一期研究,2002年至2007年完成了第二期研究,自2007年起又进行了为期4年的第三期研究,这是一个有10多个国家参加的国际合作研究项目。其代表研究文献是《电子文件真实性的永久保存:InterPARES项目的研究发现》[②]。

2000年6月,世界银行与国际文件管理联合会联合发起了包括16个国家、地区和国际组织参与的"电子时代基于凭证性的电子文件管理"[③]的国际合作项目,其关键就是要解决具有真实性的电子文件的管理问题。

2001年国际档案理事会(简称 ICA)成立了专门工作组,研究电子文件真实性问题,并向联合国教科文组织提出了研究对策与建议,2002年11月正式发布了《电子文件的真实性:向联合国教科文组织的报告》[④]。2005年4月国际档案理事会又发布了最新研究进展——《电子文件管理:档案工作者指南》[⑤]。

2005年8月,美国国家档案馆(National Archives and Records Administration,简称 NARA)正式宣布开发电子文件档案馆(简称 ERA)[⑥]系统,该系统是目前国际上影响最大的数字档案馆项目,其目标就是在任何时间、对美国任何政府部门所产生各种类型的电子文件实施安全、有效、长期的管理。

此外,如澳大利亚、加拿大、欧盟等也都在开展电子文件管理研究项目。由此可见,电子文件管理是当今世界上备受高度关注的热点研究领域,这也对我国电子

---

① The International Research on Permanent Authentic Records in Electronic Systems (InterPARES)[EB/OL]. [2009-09-04] http://www.interpares.org.

② InterPARES, The Long-term Preservation of Authentic Electronic Record: Findings of the InterPARES Project[EB/OL]. [2009-09-04] http://www.interpares.org/book/index.cfm.

③ A World Bank/International Records Management Trust Partnership Project, Evidence-Based Governance in the Electronic Age[EB/OL]. [2009-09-04] http://www.irmt.org/Images/documents/research_reports/background_information/project_prospectus/IRMT_project_prospectus.pdf.

④ the International Council on Archives Committee on Archival Legal Matters. Authenticity of Electronic Records: A Report Prepared For UNESCO [R/OL]. (November 2002)[2009-09-04] http://www.ibls.com/internet_law_news_portal_view.aspx? id=1909&s=latestnews.

⑤ ICA Study 16 - Electronic Records: A Workbook for Archivists[EB/OL]. (Apr. 2005)[2009-09-04] http://www.ica.org/en/node/30273.

⑥ Electronic Record Archives(ERA)[EB/OL]. [2009-09-04] http://www.archives.gov/era/.

文件管理[①]提出了更严峻的挑战。

1.1.3 研究问题的提出——“基于 XML 电子文件集成置标”是电子文件管理中一个急需研究的课题

提出问题是科学分析、解决问题的重要前提。系统科学给出了研究“问题”的模型并揭示出“问题”提出的原理,“问题”的辩证认识论的含义是主观与客观矛盾的概括与抽象。当研究对象确定之后,研究者与研究对象就构成了一个系统,在该系统中,也必然会存在研究主体理想认识与研究对象客体现实间的矛盾。

电子文件管理中有许多研究问题需要解决,但是其中一个关键就是基于可扩展置标语言(简称 XML[②],下同)的电子文件集成置标问题,这在当前电子文件管理领域中是一个急需研究的课题,为此本书研究对象是基于 XML 的电子文件置标,根据矛盾的特殊性,在前期实践与文献调研的基础上,可以明确本书的研究“问题”实际就是“基于 XML 电子文件置标理论与应用研究的实际状况”不能满足“基于 XML 电子文件置标理论与应用研究的科学假设的理想状况”的矛盾。

概括地说,本书研究问题的提出主要基于以下几点:

(1) 研究的必要性与重要性。电子文件管理的最终目标是其凭证性的确保,而基于 XML 的电子文件置标对该目标贯穿全生命周期的实现是至关重要的组成部分,离开了基于 XML 的电子文件置标,电子文件的凭证价值就难以实现。

(2) 研究的可行性。目前在文件/档案领域已有关于基于 XML 的电子文件置标的研究项目与实践应用,这都为基于 XML 的电子文件集成置标提供了实践上的可应用性基础,也为其研究的可行性创造了条件。

(3) 研究的紧迫性。尽管在当前文件/档案领域已经有了关于基于 XML 的电子文件置标的实践应用,但是如何实现对基于 XML 的电子文件的多维集成置标,既缺乏理论上的深入研究,又缺乏实践上的集成应用。因此,如何将基于 XML 的电子文件置标从理论上、实践上作为一个整体加以研究,使其符合确保电子文件凭证性的需求,是当前电子文件管理中迫切需要解决的核心问题。

---

① 薛匡勇. 电子文件管理研究——中国首届档案学博士论坛综述之二[J]. 湖北档案,2002(1):12-14.

② W3C. Extensible Markup Language (XML) 1.0 (Fifth Edition)(W3C Recommendation 26 November 2008)[S]. http://www.w3.org/TR/2008/REC-xml-20081126/.

1.2 研究目的、内容与前瞻性

1.2.1 研究目的

随着人类由信息时代逐渐步入知识时代，电子文件作为一种重要的战略资源、国家软实力，已成为国际文件/档案领域及世界各国关注的热点与争相抢占的制高点。因此，如何保证电子文件的凭证性，如何实现电子文件的长久保存，就成为电子文件管理中的一个带有根本性的核心问题。围绕着这个问题，本书从“基于XML的中国电子文件集成置标理论与实践”出发，进行深入的研究，其目的就是从多维集成置标的视角、从理论与应用两个层面对中国电子文件的凭证性和长久保存的实现方式作一科学的、全面的探索，从而为电子文件在其整个生命周期中的科学置标、科学管理提供理论与实践支撑。

1.2.2 研究内容和结构设计

本书的研究内容是“基于XML的中国电子文件集成置标理论与应用研究”，即以可扩展置标语言(XML)为基础，以中国电子文件为核心，对电子文件置标的理论与应用进行集成研究。

所谓的“中国电子文件”是相对于国外电子文件而言的，主要是指本书所研究的置标电子文件文本，其在逻辑类型上主要集中于国内各类电子文件文本，如中国的电子公文文本、中国电子文件的背景信息文本、中国档案机读目录文本，及中国档案主题词表文本等，即其研究范畴是中国国内电子文件。

所谓的“集成”，其哲学内涵是指由系统的整体性及系统核心的凝聚作用而导致的使若干相关部分或因素合成为一个新的统一整体的建构及其有序化过程[①]。而所谓的“集成置标”，就是指以中国电子文件为核心，从时间维度的电子文件生命周期视角出发，以确保电子文件凭证价值得到长期、永久的保存、共享与利用；从空间维度的文本视角、置标所用语言视角、人一机环境视角出发，对中国电子文件置标进行多维、综合的系统研究，即不仅要研究中国电子文件的逻辑置标、附含语义

① 李宝山，刘志伟. 集成管理——高科技时代的管理创新[M]. 北京：中国人民大学出版社，1998.